https://doi.org/10.1038/s41592-025-02759-3

Calling all data

As life sciences research becomes enmeshed in the age of AI, real experimental data are more valuable than ever.

In attending several recent scientific conferences across diverse areas of biology, members of our editorial team have all made the same observations: no matter the field, there are calls for more data. In this age of AI, the value of experimental data goes far beyond supporting the conclusions of a research paper. Particularly as advanced machine learning approaches such as foundation models and large language models are being developed and deployed at dizzying pace, researchers are facing the hard truth: they lack sufficient data to train these huge models properly.

One success story is in the protein structure prediction field, where deep learning-based methods such as AlphaFold can now determine 3D structure with high accuracy. This feat was dependent on the existence of a long-term, stable, accessible and quality-controlled experimental data resource — the Protein Data Bank — that could be used for training. Yet, while there is also great interest in predicting 3D structures of RNA molecules, researchers attempting to develop such tools are finding that the available experimental structural RNA data are just too sparse for proper training[1].

It is certainly tantalizing to consider the many ways that generalist foundation models could serve biology. It does not surprise us that many researchers are forging ahead despite the gaps in experimental data. *Nature Methods* has published several foundation models, including scGPT[2], scFoundation[3] and Nucleotide Transformer[4] for various tasks in genomics and transcriptomics; UniFMIR[5], BiomedParse[6] and µSAM[7] for tasks in image analysis; and xTrimoPGLM[8] for tasks in protein sequence analysis. However, we also do appreciate that such models do not necessarily outperform more classical models for specific tasks. It is still an open question whether these models represent the future standard or whether smaller, custom-trained models will prevail in the end.

We are also keeping our eye on some big AI projects in biology to see how these progress. For example, several institutions, including privately funded organizations such as the Howard Hughes Medical Institute and Chan Zuckerberg Initiative, are investing large sums in AI-driven science, with the latter in particular focusing on developing a 'virtual cell' model.

Some researchers are using creative computational strategies to address the data gap. One example of this is transfer learning, whereby a model pretrained on a large amount of data is augmented with a smaller amount of specialized data to improve performance for a particular application. Others are trying to augment experimental training data themselves with real data that have been manipulated in some way (rotated, inverted, and so on), or by generating realistic synthetic data. Still, when it comes to data, nothing compares to the real thing.

What can we do to better feed these data-hungry models? Both the research community and funders need to do even more to recognize the importance of data generation. Unfortunately, for many laboratories, data generation without aim to address a clear hypothesis is not viewed as 'science' but as routine work. We need to change this perception. There are notable large-scale consortium projects (such as the Human Cell Atlas and Human Tumor Atlas Network) that focus on high-quality data generation, but even individual laboratories producing relatively small-scale datasets can make valuable contributions to the greater good. As foundation models are intended to be generally applicable to various tasks, data that are collected by different laboratories, using different instruments, on different samples, and even with varying data quality are crucial. DOIs for datasets and scalable data repositories are great steps in the right direction, but we need a better reward system to encourage data sharing.

Storing and sharing data is hard. Data repositories need stable support from funders. High-quality image-based data can be colossal (even approaching the petabyte scale) and require expensive computing solutions to store and share them. In many fields, appropriate data repositories just do not exist, making it more difficult for researchers to embrace a culture of sharing. We commend the development of resources such as the BioImage Archive and CryoET Data Portal, both of which aim to provide data and metadata in standardized, machine-readable forms.

Perhaps contrary to popular belief, there are many journals that are interested in publishing large-scale data resources, and *Nature Methods* is one of them. Our Resource format is designed exactly for the purpose of presenting datasets with potential high value to a broad community[9]. The technology used to generate such data need not be novel, nor do we require novel biological insights; the main criterion we look for is that the data will be useful for a broad community. One such example is LIVECell[10], which is a high-quality dataset of phase contrast images from 1.6 million diverse cells.

New computational methods that can do more with less data and that integrate data collected using different techniques at different biological scales will continue to be important. But we also still need new experimental technologies to help to solve the data generation gap — particularly in the spatial biology realm — that can analyze samples in higher throughput at a cheaper price. Though scientists working today have access to amazingly advanced sequencers, mass spectrometers and microscopes, we still foresee experimental methods and technology development having a central part to play in the age of AI.

Published online: 10 July 2025

References

1. Kwon, D. *Nature* **639**, 1106–1108 (2025).
2. Cui, H. et al. *Nat. Methods* **21**, 1470–1480 (2024).
3. Hao, M. et al. *Nat. Methods* **21**, 1481–1491 (2024).
4. Dalla-Torre, H. et al. *Nat. Methods* **22**, 287–297 (2025).
5. Ma, C., Tan, W., He, R. & Yan, B. *Nat. Methods* **21**, 1558–1567 (2024).
6. Zhao, T. et al. *Nat. Methods* **22**, 166–176 (2025).
7. Archit, A. et al. *Nat. Methods* **22**, 579–591 (2025).
8. Chen, B. et al. *Nat. Methods* **22**, 1028–1039 (2025).
9. *Nat. Methods* **20**, 773 (2023).
10. Edlund, C. et al. *Nat. Methods* **18**, 1038–1045 (2021).